Linear model to assess the scale's validity of a test.

(Submitted to AERA Meeting 2007, Accepted for the session: "New Developments in Measurement Thinking", SIG-Rasch Measurement)

Agustín Tristán
Instituto de Evaluación e Ingeniería Avanzada, S.C.
IEIA, Mexico

Rafael Vidal
Centro Nacional de Evaluación para la Educación Superior, A.C
CENEVAL, Mexico.

A linear model that concretes the idea by Wright & Stone to assess the quality of a test is proposed based on the experience on several real tests. The model is a "test design line" distributing uniformly the items of the test centered on 0 logits. The "test design model" is related to the "mean absolute difference", a single parameter useful to determine the distribution of the items, the influence of the bias of the scale and the test width. Results and applications of the model, from test design to test analysis and calibration are shown.

Key words: Domain theory, quality, Rasch, scale, test, validity.

Validity is the most important attribute related to the quality of a test or to the quality of the decisions taken with the results of a test. The main focus of validity-centered designs is the formal set of objective characteristics of the "domain theory" (Bunderson, 2005) where the measurement scales associated with the Rasch model, have these attributes of invariance: (1) invariance of sample – measures independent of the sample of persons; (2) invariance of task – measures independent of the set of items; (3) invariance of unit and zero – measures have approximately equal intervals, a zero and the unit is constant; (4) invariance of interpretive – measures coherent with a construct framework, with milestones and level descriptors.

Other approaches are the hierarchical complexity of the tasks (Commons & Miller, 2001) and the covariance structure modeling (Raykov, 2005). Contrary to the definition of validity by Messick (1998) and Cronbach & Meehl (1955), the simple definition of causality for construct validity proposed by Boorsboom et al (2004), provides a framework for logistic models. Bond (2004) explains how the Rasch model meets the requirements by Messick.

Wright & Stone (1988) pointed out the difficulty to handle the definitions of validity, because it is not evident how to choose the data to be used in a relevant and perhaps "correct" criterion. They focus on two evidences: (a) order validity (the meaning of the calibration of the items, according to an order from easy to difficult) and (b) fit validity regarding the discrepancies between observed and expected responses (item and person measures correspond to a "useful definition of measurement").

Wright & Stone (1979) suggest the rules to obtain the "shape of a best test"; from them, some general ideas may be adopted, specifically: (a) the range of the items' difficulties (the test width must cover the full range of persons' abilities); (b) the mean of the items' difficulties (it must be as close as possible to the mean of the persons' abilities); (c) the distribution of the items (uniformly distributed in the full range of difficulties). A valid scale of a test must contain the set of properties indicated by those authors.

The questions are: a thermometer having marks unevenly distributed from 35 to 42 ^{o}C is a valid measurement instrument? Is it needed to have equally distributed intervals (for instance, every ½ ^{o}C) for the marks in a valid measurement instrument? If the instrument does not produce high quality measures and interpretations, then we face a validity problem. The two mentioned thermometers will have a different *reliability* if they have different intervals, and probably will influence the *objectivity* of the measures and conclusions produced with the results, but what can we say about the *validity of the scale*?

The recommendations by Wright & Stone and Bunderson bring the characteristics of a measurement instrument, independently of its purpose, these authors need an objective, linear and additive scale based upon the Rasch model (Bond & Fox, 2001). But the Rasch model produces mathematically a linear scale from any set of items; then why do we

Correspondence may be addressed to: ici_kalt@yahoo.com

need to follow the recommendations of these authors concerning the distribution of items? Some ideas are:

a) A uniform distribution controls the scale in logits.
b) The test measurement error can be known in advance.
c) The item's Rasch measures are evenly positioned, avoiding zones with excess of items and zones where items are missing.
d) It provides an objective way to compare between different tests or populations.

The concept of "scale validity" (SV) has different interpretations in the references, such as: (a) the valid choice of ordered categories in a rating scale (McDonald, 2004); (b) the dimensionality of the scale (Suchman, 1950), (c) the hierarchical model inherent to the scale through ordering theory (Byers & Byers, 1998) and (d) the way to evaluate the quality of the scale and the comparability of the results coming from measurement (Dawis, 1987). For O'Connor (2004), SV largely means the same as test validity; for DeVellis (1991) scale and measurement instrument are synonyms, while for Wright & Stone (2004) the scale has the meaning of the measurement construct we are intending to assess, and it also refers to the concept of the "yardstick", corresponding to the validity of the measurement scale (in logits or other units), that has to be with the fit of the data of the Rasch model[1]. Our concern in SV focus on the quality of the "yardstick scale" produced by the items; to know objectively if the item distribution in a test corresponds to a valid yardstick.

Figure 1 shows the output from Winsteps sample data using the Knox Cube Test. Items are not uniformly distributed: "stacks" and gaps are evident; Wright & Stone propose some ideas to improve the construction of the test, including an item close to 0 and avoiding the items with extreme difficulties.

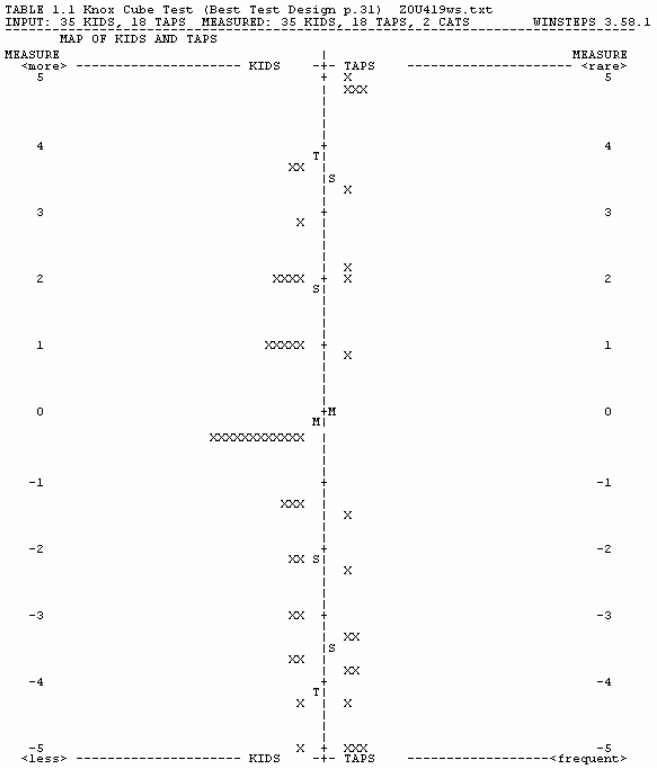

Figure 1

Table 2.1 from Winsteps (Figure 2) provides the most probable response for the test.

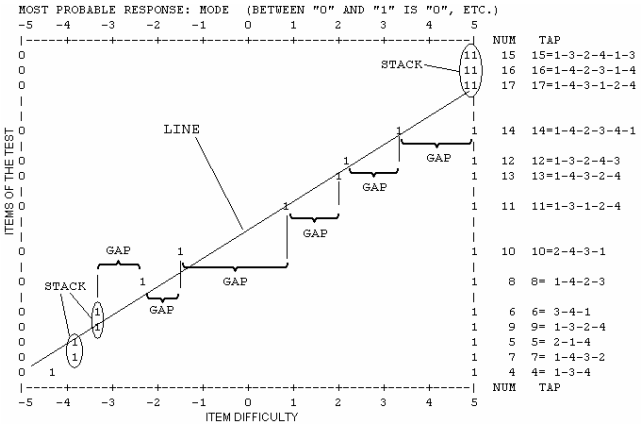

Figure 2

Wright & Stone pointed out three features in this map: (1) stacks (two or more items are located in the same measure); (2) gaps (a set of measures not covered by items); (3) line (a theoretical uniform distribution of the items' difficulties). Wright & Stone say (page 38) that "the straighter the line, the fewer the distortions and the closer the data to the line, the more uniform the conjoint relation between items and persons, and the clearer the definition of the metric of the yardstick that was built to define the variable", but they do not propose a model for this line.

[1] Linacre J.M. Personnel communication, 2006.

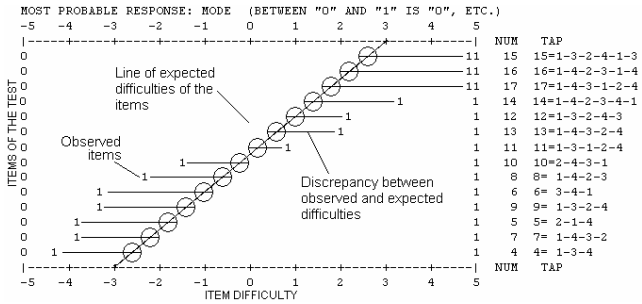

Figure 3

The discrepancies among observed and expected difficulties, can be easily calculated (Figure 3).

The construction of the line involves two issues:

a) The "test design line" definition.
b) The acceptable depart of the discrepancies observed-theoretical difficulties of the items.

Two proposals are available for the "test design line":

a) The normative line according to the mean and standard deviation of the measures of the persons (Wright & Stone, 1979).
b) The best fit line to the observed difficulties of the items (Wright & Stone, 2004).

The mean difficulty located in 0 is a good proposal, because it reduces or eliminates any bias on the design of the test. But the width cannot be normative or related to the best fit (some objections are even presented by Wright & Stone, 2004, pp. 39): as a descriptive model could not be useful for test design. Instead, a fixed position of the "test design line", permits comparisons between tests, and can be useful for design and calibration.

Linear model for the test scale.

The position of the "test design line" (TDL) has been obtained from a meta-analysis of tests from different countries and educational levels[2] (Tristan & Molgado, 2007). A width of 3 logits (from -1.5 to +1.5) provides a common range of the measures in real tests. Furthermore, the interval corresponds to the limits of p=0.82 and p=0.18, corresponding very closely to the classical values recommended by various investigators, suggesting the design of tests with item difficulties in the interval (20%-80%)[3].

The equation of the TDL[4] for N items is:

$$D=W(I-1)/(N-1)-LL \qquad [1]$$

Where:
D = Difficulty of item I (from 1 to N).
LL = Lower difficulty of the test, in logits. Suggested: LL = -1.5 logits.
W = Width of the test in logits (W = 2 x abs(LL)). Suggested: W=3.0 logits.

The test responsible may define LL and W, according to his design; but the TDL[-1.5,1.5] produced good results for design and assessment of the scale of tests applied in Mexico, El Salvador and Colombia, from preschool to graduate levels.

Items to the right of the TDL are harder than the expected difficulty (and to the left of the TDL are easier that the expected difficulty[5]). The discrepancies between observed and expected item difficulties can be calculated and the mean absolute difference for N items is:

$$MAD = [\Sigma\, Abs(D_{observed} - D_{expected})]/N \qquad [2]$$

Where:
MAD = Mean absolute difference
D = Item difficulty

For real tests a MAD above 0.25 logits (called "¼ logit rule") indicates a high discrepancy among difficulties[6]. Discrepancies may be due to: (a) the items are not uniformly distributed, (b) the mean of difficulties is far from zero or far from the persons mean, or (c) the width of the test is bigger than 3 logits. These issues may be solved through an item bank with sufficient calibrated items. The MAD is a straightforward fit parameter, a more complicated one could be proposed, but our experience indicates that a simple formula is better accepted by teachers and test designers.

---

[2] Tristan L.A. & Molgado, D. (2007) Limits of measures in tests, a meta-analysis. Internal Report. IEIA, México.

[3] The relationship between (20%,80%) is (1.38,-1.38). In this interval, p values and measures in logits are practically proportional, following the function: p = -0.2244b + 0.5, with a fit r= 0.99896.

[4] Formula [1] is equally valid for classical values, but LL and W must be defined in the same units, according to the model in percent of correct answers or fractions of unity.

[5] For Classical Test Theory, difficulty is defined with correct answers, the interpretation is contrary to this one.

[6] The ¼ logit rule for the maximal MAD represents 1/12 of the test width. For the classical interval (20%,80%), the maximal MAD corresponds to maxMAD=(80-20)/12=5% discrepancy, a very simple value to remember.

The "Test Design Line" is defined by three values: TDL[LL,UL,MAD], for example:TDL[-1.5,+1.5,0.2], means these properties of a test:

a) Lower limit, LL= -1.5, upper limit, UL = +1.5 (in logits)
b) Uniform distribution of items, MAD = 0.2< ¼ logit
c) Mean value, centered in 0 logits; M = (LL+UL)/2
d) Width equal to 3 logits; W = UL-LL
e) No presence of bias; B = -M and MAD<0.25

The theoretical design line that we are proposing is TDL[-1.5,+1.5,0]. For the Knox Cube Test previously shown, we get a TDL[-5.75,5.76,0.66] (Figure 4).

Given a TDL[LL, UL, MAD], the real difficulties of the items may be below LL or above the UL, and also the MAD may increase. Therefore a real test may include easier or harder items but keeps the uniform distribution of the items, if MAD is below 0.25.

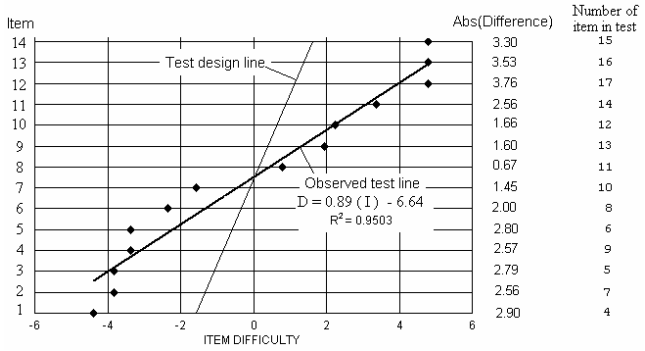

Figure 4

The TDL provides the distribution of the items in a test, and a kind of snapshot of the scale for a given test. The graphical representation, and the MAD provide an evidence of the validity of the scale.

Applications of the "test design line"

A test of professional competencies (combining abilities and knowledge) has been applied to a population of 322 persons. The test (St1) has 39 multiple-choice items, organized in three subtests of items (St2=13, St3=11 and St4=14 items). Each subtest has been designed with a computer-based item bank, distributing the items according to TDL[-1.5,+1.5,0].

Two subgroups may be found in the population: Pop1 (226 mainly competent persons), and Pop2 (96 mainly non-competent persons). The difference in the mean ability measure between both populations is 0.84 logits for the whole test (0.95 for St2, 0.66 for St3 and 1.0 for St4).

Verification of the "scale validity" for both populations.

Figure 5 compares the TDL for Pop1 and Pop2, the mean of the items' difficulties on each test has been shifted to the 0 of the measures of the persons. The test for Pop1 is closer to the TDL (MAD=0.2) than the same test for Pop2 (MAD=0.64), so the test has an acceptable measurement scale for Pop1. The difference in mean between both populations (0.84 logits) reflects the gap in abilities among them.

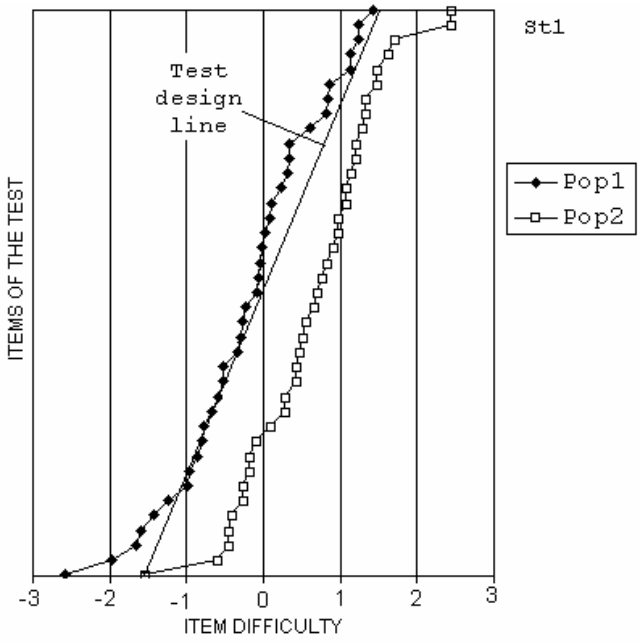

Figure 5

Other conclusions are:

a) For Pop2, the test is harder than expected.
b) For Pop1, the test is well targeted.
c) For Pop1, there are more easier items that expected: half of the items lie to the left of the TDL.

When the items difficulties are shifted to the mean, it is difficult or even impossible to compare among populations and items, then it seems better to have a fixed TDL. Figures 6a to 6c show the three subtests for both populations, with conclusions similar to those presented previously.

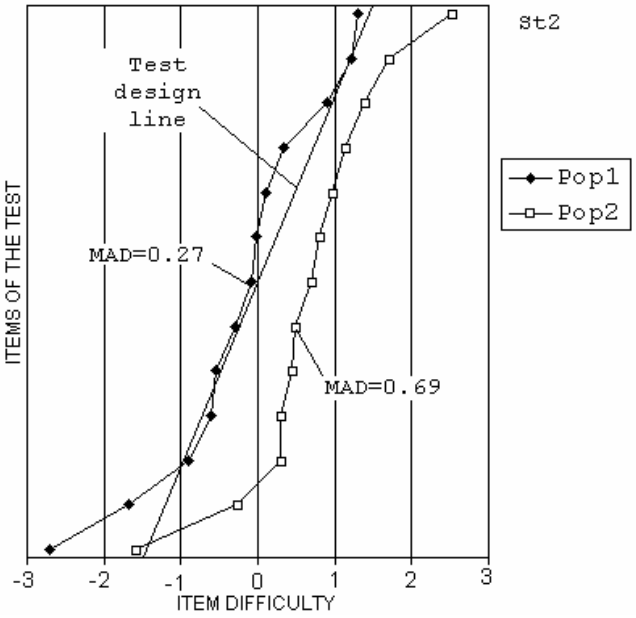

Figure 6.a

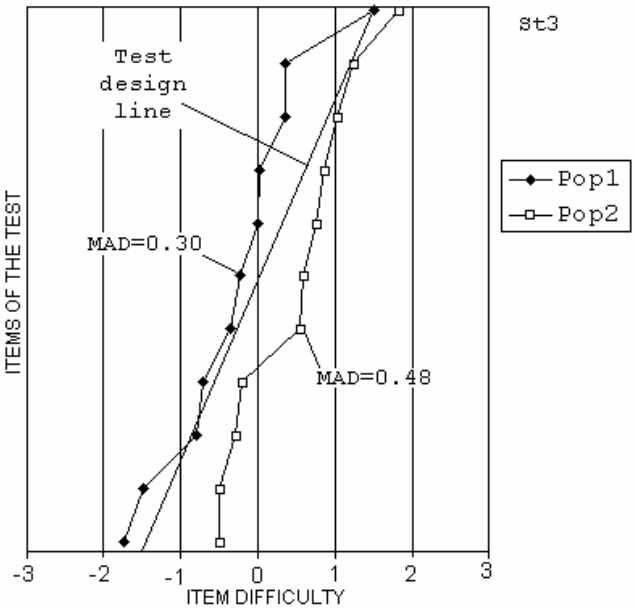

Figure 6.b

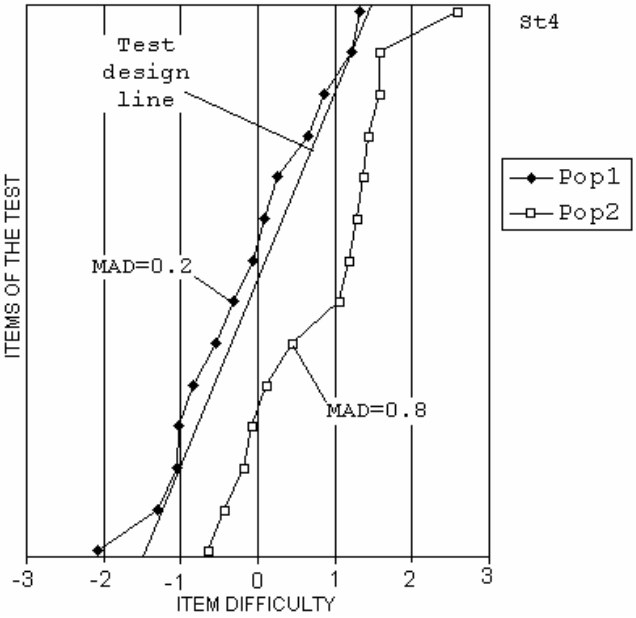

Figure 6.c

Comparisons among populations.

If the tests are centered on its mean difficulty, it is possible to check the uniformity of the distribution, against the same reference. The test and the subtests have an acceptable item distribution (MAD<0.25) and seems to be the same test independently of the populations (Figure 7).

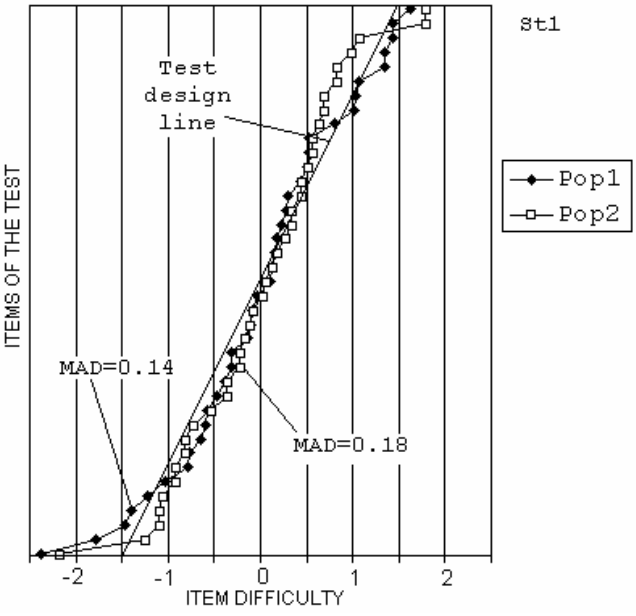

Figure 7

Comparison with the line of the best fit.

The line of the best fit can be found for the test, and it can be compared against the TDL. The "observed test line", provides the limits, width and mean of the observed test.

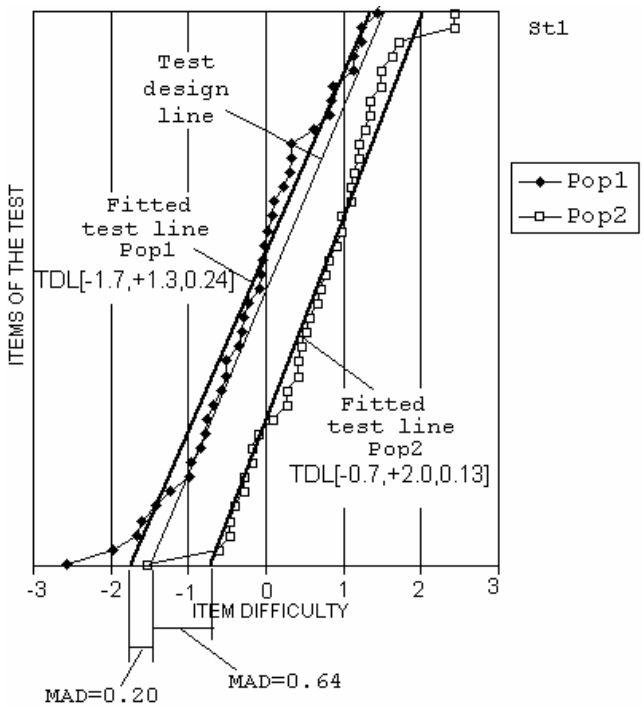

Figure 8

The fitted line test for Pop1 (Figure 8) is equivalent to a design TDL[-1.7,+1.3,0.24], very close to the reference TDL (MAD=0.20); for Pop2, the test is equivalent to a design TDL[-0.7,+2.0,0.13], (MAD=0.64), confirming the difference in measure of 0.84 among the populations.

Other applications

A unique TDL[-1.5,+1.5,0] used for the test and the subtests, permits to compare any test against a single model. A fixed TDL is not only useful for test design, it provides a tool to calculate and improve the reliability and the standard error of a "reference test" (Tristan & Vidal, 2001). It also provides a tool to set passing scores, combined with anchored and the bookmark methods. Furthermore it may help to decide the best positions to eliminate item stacks, and also to determine the effect if new items are including in the gaps of a test.

Conclusions

The concept of "scale validity" combines the attributes of invariance for a validity-centered design and the order and fit validities for the item distribution on a test. The model as a "test design line" has the following elements: (a) uniform distribution of items difficulties; (b) range of difficulties covering all the spectrum of person's abilities; (c) mean difficulty at the center of the interval; (d) no bias of the test difficulty. The mean absolute difference is a quantitative parameter that reflects the distribution of the items and the limits of the design of a given test. The model has been successfully used since 2001, to identify the "scale validity" of tests in Mexico, El Salvador and Colombia, from preschool up to professional level.

References

Bond T.G. & Fox C.M. (2001) *Applying the Rasch model*. Erlbaum, NJ Pp. 4-8.

Bond T.G. (2003) *Validity and assessment: a Rasch measurement perspective.* Metodologías de las ciencias del comportamiento 5(2), 179-194

Borsboom, D., Mellenbergh G.J. & ven Heerden J. (2004) *The concept of validity*. Psychological review, 111(4),1061-1071

Bunderson, C.V. (2005) *Developing a domain theory in fluent oral reading*, Advances in Rasch Measurement, Vol. 1. MN. JAM Press.

Byers C. & Byers W.A. (1998) *Sliding scale: a technique to optimize the assessment of knowledge level.* Int. Pers. Manag. Ass. Assessment Council. Chicago. June. 5 pp.

Commons M.L. & Miller P.A. (2001) *A quantitative behavioral model of developmental stage bases upon hierarchical complexity theory*. Behavior Analyst Today, 2(3), 222-240

Cronbach L.J. & Meehl, P.E. (1955) *Construct validity in psychological tests.* Psychological Bulletin, 52, 281-302

Dawis R.V. (1987) *Scale construction.* Journal of Counseling Psychology. 34(4), 481-489

DeVellis R.F. (1991) *Scale development.* Applied Social Research Methods Series. Vol. 26. Sage, NewBury Park. Pp. 8-11.

Linacre J.M. (2006) A user's *guide to Winsteps.* Winsteps.com

O'Connor (2004) *Measuring quality of life in health.* Elsevier Science Health Science.

McDonald J.A.L. (2004) *The optimal number of categories for numerical rating scales*. PhD Dissert. Coll. Education. Univ.of Denver. 170 pp.

Messick, S. (1998) Validity. In R.L.Linn (ed) Educational measurement (pp 13-103). Washington DC: Am. Council on Education and National Council on Measurement in Education.

Raykov T. (2005) *A method for testing group differences of scale validity in multiple population studies.* British J. of Mathematical and Statistical Psychology, 58, 173-184.

Suchman E.A. (1950) *The logic of scale construction.* Educational and Psychological Measurement. Vol. X. 79-93

Tristan L.A. & Vidal R. (2001) *Contribution to the study of error measurement* Notas sobre evaluación criterial, N.13.IEIA México, 5 pp.

Wright B.D. & Stone M.H. (1988) *Validity in Rasch measurement.* Research memorandum 54. MESA. University of Chicago. 12 pp.

Wright B.D. & Stone M.H. (1979) *Best test design.* MESA Press. Chicago.pp 133-140

Wright B.D. & Stone M.H. (2004) *Making measures*. The Phaneron Press.Chicago. USA. Pp.35-39.